STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1490–1507
Published online 20 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3026

Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data

Tsung I. Lin[1,*,†] and Jack C. Lee[2]

[1]*Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan*
[2]*Graduate Institute of Finance, National Chiao Tung University, Hsinchu 30050, Taiwan*

SUMMARY

This paper extends the classical linear mixed model by considering a multivariate skew-normal assumption for the distribution of random effects. We present an efficient hybrid ECME-NR algorithm for the computation of maximum-likelihood estimates of parameters. A score test statistic for testing the existence of skewness preference among random effects is developed. The technique for the prediction of future responses under this model is also investigated. The methodology is illustrated through an application to Framingham cholesterol data and a simulation study. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: ECME algorithm; maximum-likelihood estimation; prediction; random effects; SNLMM

1. INTRODUCTION

The normal linear mixed model (NLMM), originally introduced by Laird and Ware [1], has become the most frequently used analytic tool for longitudinal data analyses with continuous repeated measures. The specification of both random effects and the error term is automatically taken to be normal for mathematical convenience. Unfortunately, such normality assumptions are too restrictive and suffer from the lack of robustness against departure from the normality assumption. Meanwhile, it may yield invalid statistical inferences when the underlying normalities are violated.

To reduce unrealistic normality assumptions in NLMM, various authors concentrated on more flexible approaches using a broader family of distributions. Pinheiro *et al.* [2] proposed a (multivariate) t linear mixed model and showed that it would perform well in the occurrence of outliers.

*Correspondence to: Tsung I. Lin, Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan.
†E-mail: tilin@amath.nchu.edu.tw

Contract/grant sponsor: National Science Council of Taiwan; contract/grant number: NSC95-2118-M-005-001-MY2

Received 13 November 2006
Accepted 22 June 2007

Verberk and Lesaffre [3] introduced a heterogeneous linear mixed model with random effects distributed according to finite normal mixtures. They provided a simple goodness-of-fit test for the investigation of heterogeneity in random effects.

An alternative way of enhancing the justification of the normality assumption in the presence of strong skewness is *via* the Box–Cox transformation; see e.g. [4, 5]. Although such transformation-based methodology may yield reasonable empirical results, the achievement of joint normality is rarely satisfied and the transformed variables are difficult to interpret. From a practical viewpoint, one needs to seek an appropriate theoretical model from a robustification of normal theory in regulating skewness.

The multivariate skew-normal (SN) distribution was introduced by Azzalini and Dalla Valle [6], and some further attractive features as well as applications are given in Azzalini and Capitaino [7]. For this class of distributions, a number of extensions and alternative formulations have been proposed during the last decade. Sahu *et al.* [8] defined a new class of multivariate SN distributions and stated that it is very convenient to implement within a Bayesian hierarchical framework. Arellano-Valle and Genton [9] studied the family of fundamental skew normal (FUSN) distributions obtained by a convolution scheme that generalizes [8] and leads to a fairly general procedure to obtain the multivariate SN densities starting from symmetric ones. For more discussions of FUSN and other proposals of the original SN distribution such as the closed SN of [10] and the hierarchial SN of [11] along with their connections and related properties, see [12, 13].

Recently, Arellano-Valle *et al.* [14] proposed a skew-normal linear mixed model (SNLMM) based on the multivariate SN distribution introduced by Azzalini and Dalla Valle [6] and Azzalini and Capitaino [7]. They assumed that both random effects and within-subject errors are distributed as SN for the sake of completeness. To the best of the authors' experiences, however, in this setting the SN distributional assumption for within-subject errors is hard to justify even when the number of observations is sufficiently large. Moreover, it is irrational to assume that the skewness parameters for within-subject errors are identical for the parsimony of model building. As can be seen in the illustrative example of [14], such representation departs from the true realities for the fitted residuals and in turn leads to less significance for the estimated skewness parameter.

In this paper, we aim at developing additional tools for a simplified version of SNLMM, more directly linked to the work of [14], in which merely the random effects are assumed to follow multivariate SN distributions. A key feature of this model is that it can be formulated as a flexible normal-truncated normal hierarchy, which is useful for random number generation and for theoretical derivations. In view of the computational perspective, we present a hybrid of EM-type and gradient-based methods alternative to [14], which is used for accelerating the calculation of maximum-likelihood (ML) estimates with the standard errors as a by-product at convergence.

In Section 2, we describe the model, define the notations and derive a few distributional properties under the complete data framework. In Section 3, we discuss how to compute ML estimates of model parameters in an efficient manner. In Section 4, a score test statistic is obtained for testing the skewness preference among random effects. Such a test can serve as a preliminary check before fitting a complex SNLMM. Inferences for the estimation of random effects and the prediction of future values are discussed in Section 5. We illustrate these results using the famous Framingham cholesterol data in Section 6. One simulation study is presented in Section 7. Some concluding remarks are given in Section 8 and technical derivations are given in the Appendices.

DOI: 10.1002/sim

2. MODEL AND NOTATION

A p-dimensional random vector $\mathbf{Y}$ is said to have a multivariate SN distribution with a $p \times 1$ location vector $\boldsymbol{\mu} \in \mathbb{R}^p$, a $p \times p$ positive-definite dispersion matrix $\boldsymbol{\Sigma}$ and a $p \times 1$ skewness vector $\boldsymbol{\lambda} \in \mathbb{R}^p$, say $\mathbf{Y} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, if its probability density function (pdf) is

$$f(\mathbf{Y}) = 2\phi_p(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})) \tag{1}$$

where $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi(\cdot)$ is the cumulative distribution function of $\mathrm{N}(0, 1)$. Note that if $\boldsymbol{\lambda} = \mathbf{0}$, the pdf of $\mathbf{Y}$ in (1) corresponds to $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density. Some essential distributional properties with regard to (1) are referred to [6, 7, 15].

We consider a generalization of NLMM in which the random effects are assumed to follow multivariate SN distributions within the family defined in (1). The model can be written as

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{b}_i \sim \mathrm{SN}_q(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}, \boldsymbol{\lambda}) \\ \boldsymbol{\varepsilon}_i &\sim \mathrm{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{C}_i), \quad \mathbf{b}_i \perp \boldsymbol{\varepsilon}_i \qquad (i = 1, \ldots, N) \end{aligned} \tag{2}$$

where the subscript i is the subject index, $\mathbf{Y}_i$ is an n_i-dimensional random response vector, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known full-rank matrices of dimensions $n_i \times p$ and $n_i \times q$, respectively, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed effects describing the population mean, $\mathbf{b}_i$ is a $q \times 1$ vector of unobservable random effects and $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ vector of residual errors assumed to be independent of $\mathbf{b}_i$. Moreover, $\boldsymbol{\Gamma}$ is a $q \times q$ unstructured symmetric positive-definite matrix, $\mathbf{C}_i$ is an $n_i \times n_i$ scaled matrix, which is a function of a small set of autocorrelation parameters $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_g)$ and depends on i only through its dimension n_i, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$ is a $q \times 1$ vector of skewness parameters. In what follows, we reparameterize $\boldsymbol{\Gamma} = \mathbf{F}^2$ for ease of computation and theoretical derivation, where $\mathbf{F}$ is the square root of $\boldsymbol{\Gamma}$, i.e. $\boldsymbol{\Gamma}^{1/2}$, containing $q(q+1)/2$ distinct elements.

According to Proposition 1 of [14], model (2) can be hierarchically represented as

$$\begin{aligned} \mathbf{Y}_i \mid \tau_i &\sim \mathrm{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{d}_i\tau_i, \sigma^2\boldsymbol{\Psi}_i) \\ \tau_i &\sim \mathrm{TN}(0, \sigma^2; (0, \infty)) \end{aligned} \tag{3}$$

where

$$\boldsymbol{\Psi}_i = \mathbf{Z}_i\mathbf{F}(\mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}})\mathbf{F}^{\mathrm{T}}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{C}_i = \boldsymbol{\Lambda}_i - \mathbf{d}_i\mathbf{d}_i^{\mathrm{T}}$$

$$\boldsymbol{\Lambda}_i = \mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{C}_i, \quad \mathbf{d}_i = \mathbf{Z}_i\mathbf{F}\boldsymbol{\delta}, \quad \boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{\lambda}) = \frac{\boldsymbol{\lambda}}{\sqrt{1 + \sum_{j=1}^{q}\lambda_j^2}}$$

and $\mathrm{TN}(\mu, \sigma^2; (a, b))$ denotes the truncated normal distribution for $N(\mu, \sigma^2)$ lying within a truncated interval (a, b).

It follows from (3) that the density of $\mathbf{Y}_i$ is

$$f(\mathbf{Y}_i) = 2\phi_{n_i}(\mathbf{Y}_i|\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{\Lambda}_i)\Phi\left(\frac{\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\sigma\sqrt{1 + \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}}\right) \tag{4}$$

with mean $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} + \sigma\sqrt{2/\pi}\mathbf{d}_i$ and covariance matrix $\mathrm{cov}(\mathbf{Y}_i) = \sigma^2((1 - 2/\pi)\mathbf{d}_i\mathbf{d}_i^{\mathrm{T}} + \boldsymbol{\Psi}_i)$.

DOI: 10.1002/sim

From (4), it is straightforward to observe that $\mathbf{Y}_i$ are independent and marginally distributed as $\mathbf{Y}_i \sim \mathrm{SN}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{\Lambda}_i, \boldsymbol{\alpha}_i)$, where $\boldsymbol{\alpha}_i = (1 + \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)^{-1/2}\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i$. Consequently, simple algebra yields

$$\tau_i \mid \mathbf{Y}_i \sim \mathrm{TN}(\mu_{\tau_i}, \sigma^2_{\tau_i}; (0, \infty))$$

where

$$\mu_{\tau_i} = (1 + \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)^{-1}\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}), \quad \sigma^2_{\tau_i} = \sigma^2(1 + \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)^{-1} \tag{5}$$

Note that, in particular,

$$E(\tau_i \mid \mathbf{Y}_i) = \mu_{\tau_i} + \kappa_i\sigma_{\tau_i} \quad \text{and} \quad E(\tau_i^2 \mid \mathbf{Y}_i) = \mu^2_{\tau_i} + \sigma^2_{\tau_i} + \kappa_i\mu_{\tau_i}\sigma_{\tau_i} \tag{6}$$

where

$$\kappa_i = \frac{\phi(\eta_i)}{\Phi(\eta_i)} \quad \text{and} \quad \eta_i = \frac{\mu_{\tau_i}}{\sigma_{\tau_i}} = \frac{\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\sigma\sqrt{1 + \mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}} \tag{7}$$

Denoting by $\mathbf{Y}=(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ the observed data, the log-likelihood function of $\boldsymbol{\theta}=(\boldsymbol{\beta}, \sigma^2, \mathbf{F}, \boldsymbol{\rho}, \boldsymbol{\lambda})$ is

$$\ell(\boldsymbol{\theta}|\mathbf{Y}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{\Lambda}_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i + \sum_{i=1}^{N}\log\Phi(\eta_i) \tag{8}$$

where $n = \sum_{i=1}^{N} n_i$ denotes the total number of observations, $\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_i(\mathbf{F}, \boldsymbol{\rho})$ and $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}$. Explicit expressions for the score vector $\mathbf{s}_{\boldsymbol{\theta}}$ and the observed information matrix $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ are derived in Appendix A.

3. MAXIMUM-LIKELIHOOD ESTIMATION

Since the ML estimators of parameters in model (2) cannot be written in closed forms, the popular Newton–Raphson (NR) algorithm can be applied to calculate or approximate ML estimates iteratively. Starting from an initial point $\hat{\boldsymbol{\theta}}^{(0)}$, the NR procedure proceeds according to

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} + \hat{\mathbf{I}}^{-1(k)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{\mathbf{s}}^{(k)}_{\boldsymbol{\theta}} \tag{9}$$

where $\hat{\mathbf{s}}^{(k)}_{\boldsymbol{\theta}}$ and $\hat{\mathbf{I}}^{(k)}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ are the score vector and the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}^{(k)}$, respectively. Other related iterative algorithms such as steepest ascent, Davidon–Fletcher–Powell, conjugate gradient, quasi-Newton and Marquardt–Levenberg, see e.g. [16], which are gradient methods such as the NR algorithm, can lead to equivalent calculations.

An oft-voiced complaint of the NR algorithm is that it may not converge unless good starting values are used. The EM algorithm [17], which takes advantage of being insensitive to the starting values as a powerful computational tool that requires the construction of unobserved data, has been well developed and has become a broadly applicable approach to the iterative computation of ML estimates. One of the major reasons for the popularity of the EM algorithm is that the

DOI: 10.1002/sim

M-step involves only complete data ML estimation, which is often computationally simple. Moreover, the EM algorithm is stable and straightforward to implement since the iterations converge monotonically and no second derivatives are required.

When the M-step of EM turns out to be analytically intractable, it can be replaced with a sequence of conditional maximization (CM) steps. Such modification is referred to as the ECM algorithm [18]. The ECME algorithm [19], a faster extension of EM and ECM, is obtained by maximizing the constrained Q-function (the expected complete data function) with some CM steps that maximizes the corresponding constrained actual likelihood function, called the CML steps. However, computation to the solution of the CML step may involve a high-dimensional search problem for which the likelihood is not guaranteed to increase. In practice, a modified NR maximization procedure such as the step-halving method can be used to guarantee that each step increases the likelihood.

Let $\boldsymbol{\tau}=(\tau_1,\ldots,\tau_N)$ be the latent vector, and let $\hat{\boldsymbol{\theta}}^{(k)}=(\hat{\boldsymbol{\beta}}^{(k)},\hat{\sigma}^{2(k)},\hat{\boldsymbol{\omega}}^{(k)})$, where $\hat{\boldsymbol{\omega}}^{(k)}=(\text{vech}(\hat{\mathbf{F}}^{(k)}),\hat{\boldsymbol{\rho}}^{(k)},\hat{\boldsymbol{\lambda}}^{(k)})$, denote the estimates of $\boldsymbol{\theta}$ at the kth iteration. Given the hierarchical representation (3), the complete data log likelihood is

$$\ell_c(\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\tau})=-\frac{n+N}{2}\log(\sigma^2)-\frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{\Psi}_i|$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\{(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{d}_i\tau_i)^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{d}_i\tau_i)+\tau_i^2\} \tag{10}$$

Given the current estimate $\boldsymbol{\theta}^{(k)}$, the E-step calculates $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})=E(\ell_c(\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\tau})|\mathbf{Y},\boldsymbol{\theta}^{(k)})$. From (10), calculation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ requires the expressions of $\hat{\tau}_i^{(k)}=E(\tau_i|\mathbf{Y},\boldsymbol{\theta}^{(k)})$ and $\hat{u}_i^{(k)}=E(\tau_i^2|\mathbf{Y},\boldsymbol{\theta}^{(k)})$, which are readily evaluated according to (6). That is,

$$\hat{\tau}_i^{(k)}=\hat{\mu}_{\tau_i}^{(k)}+\hat{\kappa}_i^{(k)}\hat{\sigma}_{\tau_i}^{(k)}\quad\text{and}\quad\hat{u}_i^{(k)}=\hat{\mu}_{\tau_i}^{2(k)}+\hat{\sigma}_{\tau_i}^{2(k)}+\hat{\kappa}_i^{(k)}\hat{\mu}_{\tau_i}^{(k)}\hat{\sigma}_{\tau_i}^{(k)} \tag{11}$$

where $\hat{\mu}_{\tau_i}^{(k)}$ and $\hat{\sigma}_{\tau_i}^{(k)}$ in (11) are μ_{τ_i} and σ_{τ_i} in (5) evaluated at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k)}$, and $\hat{\kappa}_i^{(k)}=\phi(\hat{\eta}_i^{(k)})/\Phi(\hat{\eta}_i^{(k)})$ with $\hat{\eta}_i^{(k)}=\hat{\mu}_{\tau_i}^{(k)}/\hat{\sigma}_{\tau_i}^{(k)}$. The CM steps then conditionally maximize $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$, obtaining a new estimate $\hat{\boldsymbol{\theta}}^{(k+1)}$. To update $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\sigma}^{2(k)}$, the closed-form maximizers for $\hat{\boldsymbol{\beta}}^{(k+1)}$ and $\hat{\sigma}^{2(k+1)}$ are given by

$$\hat{\boldsymbol{\beta}}_i^{(k+1)}=\left(\sum_{i=1}^{N}\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\Psi}}_i^{-1(k)}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\Psi}}_i^{-1(k)}(\mathbf{Y}_i-\hat{\mathbf{d}}_i^{(k)}\hat{\tau}_i^{(k)})$$

$$\hat{\sigma}^{2(k+1)}=\frac{1}{n+N}\left(\sum_{i=1}^{N}\hat{\mathbf{e}}_i^{(k+1)^{\mathrm{T}}}\boldsymbol{\Psi}_i^{-1(k)}\hat{\mathbf{e}}_i^{(k+1)}-2\sum_{i=1}^{N}\hat{\tau}_i^{(k)}\hat{\mathbf{d}}_i^{(k)^{\mathrm{T}}}\hat{\boldsymbol{\Psi}}_i^{-1(k)}\hat{\mathbf{e}}_i^{(k+1)}\right.$$

$$\left.+\sum_{i=1}^{N}\hat{u}_i^{(k)}(1+\hat{\mathbf{d}}_i^{(k)^{\mathrm{T}}}\hat{\boldsymbol{\Psi}}_i^{-1(k)}\hat{\mathbf{d}}_i^{(k)})\right)$$

where $\hat{\mathbf{e}}_i^{(k+1)}=\mathbf{Y}_i-\mathbf{X}_i\hat{\boldsymbol{\beta}}_i^{(k+1)}$, $\hat{\mathbf{d}}_i^{(k)}=\mathbf{Z}_i\hat{\mathbf{F}}^{(k)}\hat{\boldsymbol{\delta}}^{(k)}$ and $\hat{\boldsymbol{\Psi}}_i^{(k)}=\mathbf{Z}_i\hat{\mathbf{F}}^{2(k)}\mathbf{Z}_i^{\mathrm{T}}+\mathbf{C}_i(\hat{\boldsymbol{\rho}}^{(k)})-\hat{\mathbf{d}}_i^{(k)}\hat{\mathbf{d}}_i^{(k)^{\mathrm{T}}}$.

DOI: 10.1002/sim

Note that the maximization of $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\omega}=(\mathbf{F}, \boldsymbol{\rho}, \boldsymbol{\lambda})$ does not have analytical solutions. We then perform a CML step by maximizing the constrained actual likelihood function with respect to $\boldsymbol{\omega}$ using the following one-cycle NR procedure:

$$\hat{\boldsymbol{\omega}}^{(k+1)}=\hat{\boldsymbol{\omega}}^{(k)}+r^{(k)}\hat{\mathbf{I}}^{-1(k)}_{\boldsymbol{\omega}\boldsymbol{\omega}}\hat{\mathbf{s}}^{(k)}_{\boldsymbol{\omega}}$$

where $\hat{\mathbf{s}}^{(k)}_{\boldsymbol{\omega}}$ and $\hat{\mathbf{I}}^{(k)}_{\boldsymbol{\omega}\boldsymbol{\omega}}$ are evaluated by using the current estimates, and the scaling number $r^{(k)}$, $0<r^{(k)}\leqslant 1$, is chosen at each iteration such that the actual likelihood function is non-decreasing in k.

The iterations are repeated until a suitable convergence rule is satisfied, e.g. $\|\hat{\boldsymbol{\theta}}^{(k+1)}-\hat{\boldsymbol{\theta}}^{(k)}\|$ is sufficiently small. Unfortunately, there are instances in which this method is notoriously slow to converge. To accelerate the convergence, a faster hybrid ECME-NR algorithm is performed by running a moderate number of ECME iterations from a poor starting value until arriving in a near neighborhood of the ML estimates, and then switching to the NR algorithm (9) to speed up the convergence. Under some regularity conditions, the asymptotic variance–covariance estimates can be approximated by plugging the ML estimate $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\mathbf{F}}, \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\lambda}})$ into the inverse of the observed information matrix.

4. A SCORE TEST FOR THE SKEWNESS PARAMETER

To assess whether the skewness parameter $\boldsymbol{\lambda}$ is significantly different from zero, we use the score test for testing the null hypothesis $H_0 : \boldsymbol{\lambda}=\mathbf{0}$ against the alternative $H_1 : \boldsymbol{\lambda}\neq\mathbf{0}$. The principal advantage of using the score test in comparison with other competitive testing procedures such as the likelihood ratio or Wald tests is that it does not require the more complex model to be fitted. To obtain the score test statistic, we first calculate ML estimates under the null hypothesis, denoted by $\hat{\boldsymbol{\theta}}_0$, and then evaluate the score vector and observed information at $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0$, denoted by $[\mathbf{s}_{\boldsymbol{\theta}}]_{\hat{\boldsymbol{\theta}}_0}$ and $[\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}]_{\hat{\boldsymbol{\theta}}_0}$, respectively. Note that $\left[\mathbf{s}_{\boldsymbol{\theta}}\right]_{\hat{\boldsymbol{\theta}}_0}$ has all components equal to 0 except for the derivative with respect to $\boldsymbol{\lambda}$.

Let $\boldsymbol{\zeta}=(\boldsymbol{\beta}, \sigma^2, \text{vech}(\mathbf{F}), \boldsymbol{\rho})$, so that $\boldsymbol{\theta}=(\boldsymbol{\zeta}, \boldsymbol{\lambda})$. The score vector and observed information matrix can be reexpressed as

$$\mathbf{s}_{\boldsymbol{\theta}}=(\mathbf{s}^{\mathrm{T}}_{\boldsymbol{\zeta}}, \mathbf{s}^{\mathrm{T}}_{\boldsymbol{\lambda}})^{\mathrm{T}} \quad \text{and} \quad \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}=\begin{bmatrix}\mathbf{I}_{\boldsymbol{\zeta}\boldsymbol{\zeta}} & \mathbf{I}_{\boldsymbol{\zeta}\boldsymbol{\lambda}}\\ \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\zeta}} & \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}\end{bmatrix}$$

respectively. The score test statistic U_s is

$$U_s=[\mathbf{s}_{\boldsymbol{\theta}}]^{\mathrm{T}}_{\hat{\boldsymbol{\theta}}_0}[\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}]^{-1}_{\hat{\boldsymbol{\theta}}_0}[\mathbf{s}_{\boldsymbol{\theta}}]_{\hat{\boldsymbol{\theta}}_0}=[\mathbf{s}_{\boldsymbol{\lambda}}]^{\mathrm{T}}_{\hat{\boldsymbol{\theta}}_0}[\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}\cdot\boldsymbol{\zeta}}]^{-1}_{\hat{\boldsymbol{\theta}}_0}[\mathbf{s}_{\boldsymbol{\lambda}}]_{\hat{\boldsymbol{\theta}}_0} \tag{12}$$

where $\mathbf{s}_{\boldsymbol{\lambda}}=(\sum_{i=1}^{N}\kappa_i\partial\eta_i/\partial\lambda_1, \ldots, \sum_{i=1}^{N}\kappa_i\partial\eta_i/\partial\lambda_q)^{\mathrm{T}}$ and $\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}\cdot\boldsymbol{\zeta}}=\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}-\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\zeta}}\mathbf{I}^{-1}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\mathbf{I}_{\boldsymbol{\zeta}\boldsymbol{\lambda}}$. Explicit expressions for $\mathbf{s}_{\boldsymbol{\lambda}}$ and $\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}\cdot\boldsymbol{\zeta}}$ are shown in Appendix A.

Under H_0, U_s is asymptotically a chi-squared random variable with q degrees of freedom. A disadvantage of the score test statistic is that it tends to give too few significant results since its asymptotic distribution is approached more slowly than that of the likelihood ratio test statistic in small samples. This is due to the fact that the score test is a first-order approximation to the likelihood ratio test. For a practical perspective, we suggest that one performs the likelihood

DOI: 10.1002/sim

ratio test to yield more accurate results when the score test gives a weak indication that the null hypothesis is inappropriate, say at the 10 per cent significance level.

5. INFERENCE FOR RANDOM EFFECTS AND PREDICTION

We consider an empirical Bayes inference for the random effects that is useful for evaluating subject-specific quantities of interest such as individual intercepts and slopes. From model (2), it implies that $\mathbf{Y}_i \mid \mathbf{b}_i \sim \mathrm{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{b}_i, \sigma^2\mathbf{C}_i)$ and $\mathbf{b}_i \sim \mathrm{SN}_q(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}, \boldsymbol{\lambda})$. The conditional density of $\mathbf{b}_i$ given $\mathbf{Y}_i$ is

$$f(\mathbf{b}_i|\mathbf{Y}_i) = \frac{f(\mathbf{Y}_i|\mathbf{b}_i)\,f(\mathbf{b}_i)}{\int f(\mathbf{Y}_i|\mathbf{b}_i)\,f(\mathbf{b}_i)\,\mathrm{d}\mathbf{b}_i} = \phi_q(\boldsymbol{\mu}_{\mathbf{b}_i|\mathbf{Y}_i}, \boldsymbol{\Sigma}_{\mathbf{b}_i|\mathbf{Y}_i})\Phi(\boldsymbol{\lambda}_b^{\mathrm{T}}\mathbf{b}_i)$$

where $\boldsymbol{\mu}_{\mathbf{b}_i|\mathbf{Y}_i} = \boldsymbol{\Gamma}\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})$, $\boldsymbol{\Sigma}_{\mathbf{b}_i|\mathbf{Y}_i} = \sigma^2(\boldsymbol{\Gamma}^{-1} + \mathbf{Z}_i^{\mathrm{T}}\mathbf{C}_i^{-1}\mathbf{Z}_i)^{-1}$ and $\boldsymbol{\lambda}_b = \sigma^{-1}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\lambda}$.

After some algebraic manipulations, the minimum mean-squared error (MSE) estimator of $\mathbf{b}_i$ obtained by the conditional mean of $\mathbf{b}_i$ given $\mathbf{Y}_i$ is given by

$$\hat{\mathbf{b}}_i(\boldsymbol{\theta}) = \Phi(\Delta_i)\boldsymbol{\mu}_{\mathbf{b}_i|\mathbf{Y}_i} + \frac{\phi(\Delta_i)}{\sqrt{1+\boldsymbol{\lambda}_b^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{b}_i|\mathbf{Y}_i}\boldsymbol{\lambda}_b}}\boldsymbol{\Sigma}_{\mathbf{b}_i|\mathbf{Y}_i}\boldsymbol{\lambda}_b \tag{13}$$

where $\Delta_i = (1+\boldsymbol{\lambda}_b^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{b}_i|\mathbf{Y}_i}\boldsymbol{\lambda}_b)^{-1/2}\boldsymbol{\lambda}_b^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{b}_i|\mathbf{Y}_i}$.

In practice, the empirical Bayes estimator of $\mathbf{b}_i$, $\hat{\mathbf{b}}_i$, can be obtained by substituting the ML estimate $\hat{\boldsymbol{\theta}}$ into (13). As a consequence, it leads to $\hat{\mathbf{b}}_i = \hat{\mathbf{b}}_i(\hat{\boldsymbol{\theta}})$. We mention in passing that the detailed expression for the error covariance matrix of $\hat{\mathbf{b}}_i(\boldsymbol{\theta})$ is somewhat complicated and hence is not shown here. Furthermore, we are interested in the prediction of $\mathbf{y}_i$, a future $v \times 1$ vector of measurements of $\mathbf{Y}_i$, given the observed measurements $\mathbf{Y}=(\mathbf{Y}_{(i)}^{\mathrm{T}}, \mathbf{Y}_i^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{Y}_{(i)}=(\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_{i-1}^{\mathrm{T}}, \mathbf{Y}_{i+1}^{\mathrm{T}}, \ldots, \mathbf{Y}_N^{\mathrm{T}})^{\mathrm{T}}$.

Let $\mathbf{x}_i$ and $\mathbf{z}_i$ denote $q \times m_1$ and $q \times m_2$ matrices of prediction regressors corresponding to $\mathbf{y}_i$. Consequently, we assume that

$$\begin{bmatrix}\mathbf{Y}_i \\ \mathbf{y}_i\end{bmatrix} \sim \mathrm{SN}_{n_i+v}(\mathbf{X}_i^*\boldsymbol{\beta}, \boldsymbol{\Omega}_i, \boldsymbol{\alpha}_i^*)$$

where $\mathbf{X}_i^* = (\mathbf{X}_i^{\mathrm{T}},\ \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{Z}_i^* = (\mathbf{Z}_i^{\mathrm{T}},\ \mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Omega}_i = \sigma^2(\mathbf{Z}_i^*\boldsymbol{\Gamma}\mathbf{Z}_i^{*\mathrm{T}} + \mathbf{C}_i^*)$ and $\boldsymbol{\alpha}_i^* = (1+\mathbf{d}_i^{*\mathrm{T}}\boldsymbol{\Psi}_i^{*-1}\mathbf{d}_i^*)^{-1/2}$ $\boldsymbol{\Lambda}_i^{*1/2}\boldsymbol{\Psi}_i^{*-1}\mathbf{d}_i^*$. Moreover, we note that $\mathbf{d}_i^* = \mathbf{Z}_i^*\mathbf{F}\boldsymbol{\delta}$, $\boldsymbol{\Lambda}_i^* = \mathbf{Z}_i^*\boldsymbol{\Gamma}\mathbf{Z}_i^{*\mathrm{T}} + \mathbf{C}_i^*$, $\boldsymbol{\Psi}_i^* = \boldsymbol{\Lambda}_i^* - \mathbf{d}_i^*\mathbf{d}_i^{*\mathrm{T}}$ and $\mathbf{C}_i^* = \mathbf{C}_i^*(\boldsymbol{\rho})$ is an $(n_i+v)\times(n_i+v)$ expanded autocorrelation matrix. Let $\boldsymbol{\upsilon}_i = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\alpha}_i^*$ and let $\boldsymbol{\upsilon}_i$, $\mathbf{C}_i^*$ and $\boldsymbol{\Omega}_i$ be partitioned conformably with $\mathbf{Y}_i^* = (\mathbf{Y}_i^{\mathrm{T}},\ \mathbf{y}_i^{\mathrm{T}})^{\mathrm{T}}$. Then,

$$\begin{aligned}
&\boldsymbol{\upsilon}_i = (\boldsymbol{\upsilon}_i^{(1)\mathrm{T}}, \boldsymbol{\upsilon}_i^{(2)\mathrm{T}})^{\mathrm{T}}, \quad \mathbf{C}_i^* = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix} \\
&\boldsymbol{\Omega}_i = \begin{bmatrix}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22}\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{C}_{11} & \mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{z}_i^{\mathrm{T}} + \mathbf{C}_{12} \\ \mathbf{z}_i\boldsymbol{\Gamma}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{C}_{21} & \mathbf{z}_i\boldsymbol{\Gamma}\mathbf{z}_i^{\mathrm{T}} + \mathbf{C}_{22}\end{bmatrix}
\end{aligned} \tag{14}$$

DOI: 10.1002/sim

Here the indices for the partitioned matrices in (14) are omitted for notational simplicity. Also, we note that $\mathbf{C}_{11}=\mathbf{C}_i$, $\mathbf{C}_{12}=\mathbf{C}_{21}^{\mathrm{T}}$, $\boldsymbol{\Omega}_{11}=\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Omega}_{12}=\boldsymbol{\Omega}_{21}^{\mathrm{T}}$. Making use of the fact that

$$f(\mathbf{y}_i|\mathbf{Y}_i,\boldsymbol{\theta}) \propto \int f(\mathbf{y}_i|\mathbf{Y}_i,\tau_i,\boldsymbol{\theta}) f(\mathbf{Y}_i|\tau_i,\boldsymbol{\theta}) f(\tau_i|\boldsymbol{\theta})\,\mathrm{d}\tau_i$$

the conditional distribution of $\mathbf{y}_i$ given $\mathbf{Y}_i$ is

$$f(\mathbf{y}_i|\mathbf{Y}_i,\boldsymbol{\theta}) = \phi_v(\boldsymbol{\mu}_{2\cdot1},\boldsymbol{\Omega}_{22\cdot1}) \frac{\Phi(\boldsymbol{\upsilon}_i^{\mathrm{T}}(\mathbf{Y}_i^*-\mathbf{X}_i^*\boldsymbol{\beta}))}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))} \tag{15}$$

where

$$\boldsymbol{\mu}_{2\cdot1}=\mathbf{x}_i\boldsymbol{\beta}+\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}), \quad \boldsymbol{\Omega}_{22\cdot1}=\boldsymbol{\Omega}_{22}-\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}$$

$$\tilde{\boldsymbol{\upsilon}}_i=\frac{\boldsymbol{\upsilon}_i^{(1)}+\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\upsilon}_i^{(2)}}{\sqrt{1+\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}}$$

The minimum MSE predictor of $\mathbf{y}_i$ is the conditional expectation of $\mathbf{y}_i$ given $\mathbf{Y}_i$, i.e.

$$\hat{\mathbf{y}}_i(\boldsymbol{\theta})=E(\mathbf{y}_i|\mathbf{Y}_i,\boldsymbol{\theta})=\boldsymbol{\mu}_{2\cdot1}+\frac{\phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}\frac{\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}{\sqrt{1+\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}} \tag{16}$$

Meanwhile, the MSE covariance matrix of predictor (16) is given by

$$E[(\hat{\mathbf{y}}_i(\boldsymbol{\theta})-\mathbf{y}_i)(\hat{\mathbf{y}}_i(\boldsymbol{\theta})-\mathbf{y}_i)^{\mathrm{T}}]=E[\mathrm{cov}(\mathbf{y}_i|\mathbf{Y}_i)]$$

where

$$\mathrm{cov}(\mathbf{y}_i|\mathbf{Y}_i)=\boldsymbol{\Omega}_{22\cdot1}-\left\{\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})+\frac{\phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}\right\} \times \frac{\phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}\frac{\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}}{1+\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}} \tag{17}$$

Notice that expression (17) does not account for errors due to the estimation of unknown parameters and hence it tends to underestimate the true error covariance matrix. The prediction of $\mathbf{y}_i$ can be obtained by substituting the ML estimate $\hat{\boldsymbol{\theta}}$ into (16), leading to $\hat{\mathbf{y}}_i=\hat{\mathbf{y}}_i(\hat{\boldsymbol{\theta}})$. Proofs of (16) and (17) are sketched in Appendix B.

6. AN ILLUSTRATIVE EXAMPLE

The Framingham heart study has examined the role of serum cholesterol as a risk factor for the evolution of cardiovascular disease. Zhang and Davidian [20] proposed a semi-parametric approach to analyze a subset of the Framingham cholesterol data, which consist of gender, baseline age and

DOI: 10.1002/sim

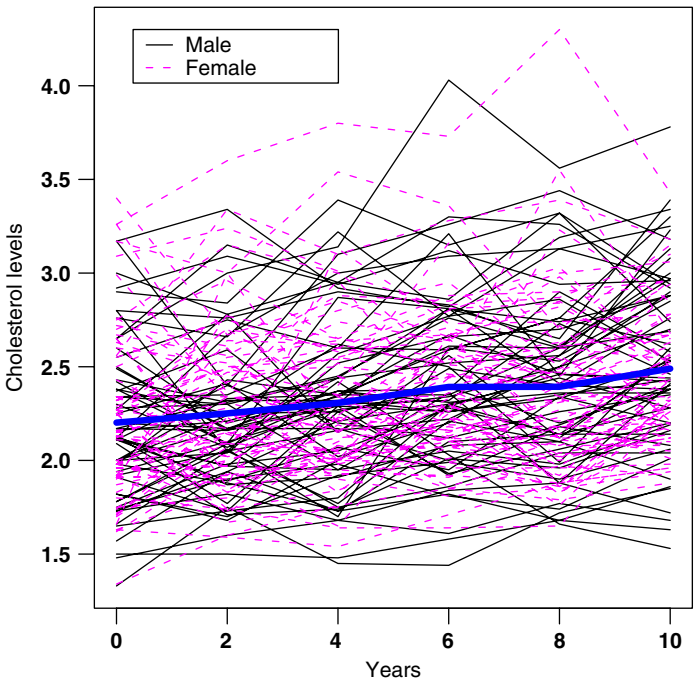

Figure 1. Trajectories of cholesterol levels for 133 participants. The thicker solid line indicates the mean profile in the study.

cholesterol levels measured at the beginning of the study and then every 2 years over 10 years, for 200 randomly selected participants. Arellano-Valle *et al.* [14] analyzed the same data set by fitting a linear mixed-effects model to it with both random effects and within-subject errors following SN distributions. Both of them come to the same conclusion that the distribution of subject-specific intercepts is non-normal.

In this section, we revisit the Framingham cholesterol data with the aim of providing additional inferences for the use of SNLMM that are not considered in [14]. To simplify the illustration, we select 133 participants (60 men and 73 women) whose trajectories of cholesterol levels as well as covariates of interest are completely observed at the duration of follow-up time. Figure 1 displays the cholesterol profiles for the 133 participants and suggests that the underlying trend seems to be linear with time and increases with a slowly moving speed.

Assuming a linear growth model with subject-specific random intercepts and slopes, we fit a linear mixed model to the data as specified by [20]

$$Y_{ij} = \beta_0 + \beta_1 \,\mathrm{sex}_i + \beta_2 \,\mathrm{age}_i + \beta_3 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, 133, \quad j = 1, \ldots, 6 \qquad (18)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ are the fixed effects of explanatory variables and the random effects $\mathbf{b}_i = (b_{0i}, b_{1i})^{\mathrm{T}}$ are assumed to have a bivariate SN distribution $\mathrm{SN}_2(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}, \boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^{\mathrm{T}}$. Moreover, the error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i6})^{\mathrm{T}}$ are assumed identically and independently $N_6(0, \sigma^2\mathbf{I}_6)$ since the ordinary least-squared residuals from the individual fits exhibit no particular pattern

DOI: 10.1002/sim

Table I. ML estimation results for model (19) with the associated standard errors in parentheses, where $\mathbf{F}$ is the square root of $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma} = \mathbf{F}^2$.

$\hat{\boldsymbol{\beta}}$	$\hat{\mathbf{F}}$		$\hat{\sigma}$	$\ell(\hat{\boldsymbol{\theta}})$	Number of parameters	AIC	BIC
1.5740	1.6869	0.7068	0.2079	−100.89	8	217.78	240.90
(0.1713)	(0.1253)	(0.1238)	(0.0104)				
−0.0338		0.9403					
(0.0619)		(0.1936)					
0.01860							
(0.0040)							
0.2787							
(0.0274)							

over time. For numerical stability as mentioned in [20], Y_{ij} are the cholesterol level divided by 100 at the jth time point for the ith participant, sex_i is a gender indicator (0 = female, 1 = male), age_i is the ith participant's age at baseline and t_{ij} is taken as $(\texttt{time} - 5)/10$, with time measured in years from the baseline.

To test for the existence of skewness preference for random effects, we start by fitting an ordinary NLMM as follows:

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{b}_i \sim N_2(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}) \\ &\boldsymbol{\varepsilon}_i \sim N_6(\mathbf{0}, \sigma^2\mathbf{I}_6), \quad \mathbf{b}_i \perp \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, 133 \end{aligned} \tag{19}$$

The resulting ML estimates, together with the maximized log-likelihood value $\ell(\hat{\boldsymbol{\theta}})$, and two penalized likelihood information criteria, AIC and BIC values (AIC = $-2\ell(\hat{\boldsymbol{\theta}}) + 2m$; BIC = $-2\ell(\hat{\boldsymbol{\theta}}) + m\log(N)$), where m is the number of parameters, corresponding to the fitted NLMM (19) are listed in Table I. The score test statistic for testing evidence of skewness among random effects gave a value of 12.17, which is highly significant compared with the χ^2_2 distribution. The score test suggests that the distribution of random effects is probably skewed. Figure 2 depicts histograms and corresponding normal quantile plots of the empirical Bayes estimates of $\mathbf{b}_i$, $\hat{\mathbf{b}}_i = \hat{\boldsymbol{\Gamma}}\mathbf{Z}_i^{\mathrm{T}}(\mathbf{Z}_i\hat{\boldsymbol{\Gamma}}\mathbf{Z}_i^{\mathrm{T}} + \mathbf{I}_6)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$, and shows that the subject-specific intercepts are positively skewed. Conversely, there are no apparent non-normal patterns for subject-specific slopes.

We thus consider an alternate model

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{b}_i \sim \mathrm{SN}_2(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}, \boldsymbol{\lambda}) \\ &\boldsymbol{\varepsilon}_i \sim N_6(\mathbf{0}, \sigma^2\mathbf{I}_6), \quad \mathbf{b}_i \perp \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, 133 \end{aligned} \tag{20}$$

The summarized ML results corresponding to the fitted SNLMM (20) are shown in Table II. All estimates are significant (compared with their standard errors) with the exception of λ_2. The results reveal that the joint distribution of random effects may, to some extent, depart from normality. Accordingly, the fitted model (20) produces smaller AIC and BIC values than those of model (19), indicating that model (20) for characterizing a more plausible assumption on the random effects is our preferred choice.

After fitting model (19) to the Framingham cholesterol data, the individual random effects, $\hat{\mathbf{b}}_i = (b_{i1}, b_{i2}), i = 1, \ldots, 133$, can be estimated by using (13) as stated in Section 5. Figure 3

DOI: 10.1002/sim

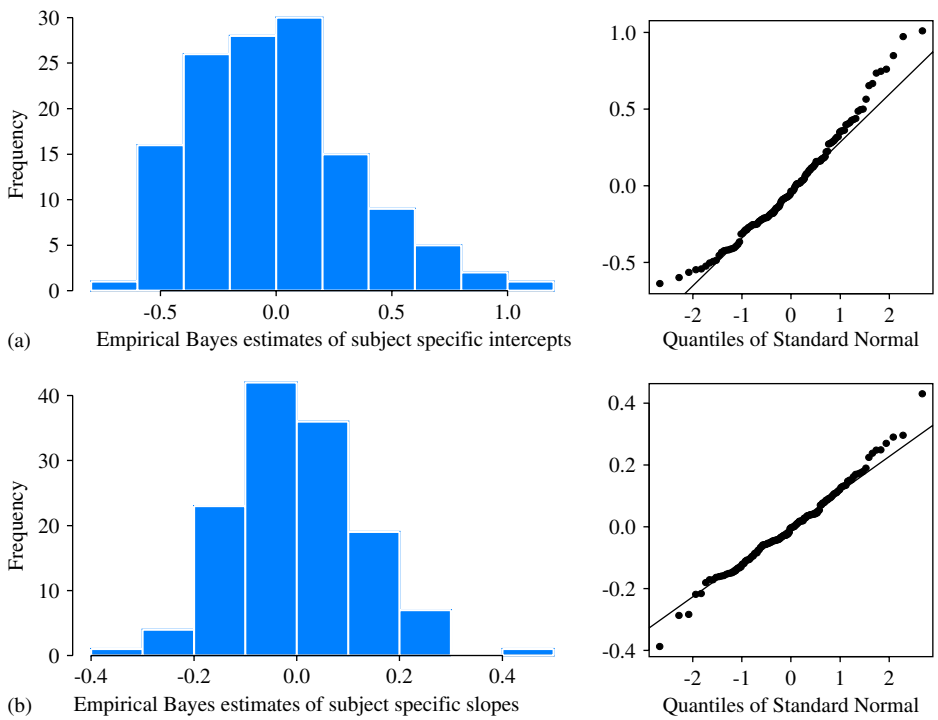

Figure 2. Histogram and normal Q–Q plots of empirical Bayes estimates of: (a) subject-specific intercepts and (b) subject-specific slopes.

Table II. ML estimation results for model (20) with the associated standard errors in parentheses, where $\mathbf{F}$ is the square root of $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma} = \mathbf{F}^2$.

$\hat{\boldsymbol{\beta}}$	$\hat{\mathbf{F}}$		$\hat{\sigma}$	$\hat{\boldsymbol{\lambda}}$	$\ell(\hat{\boldsymbol{\theta}})$	Number of parameters	AIC	BIC
1.2898	2.6799	0.2752	0.2077	4.7052	−94.18	10	208.36	237.26
(0.1731)	(0.1977)	(0.0883)	(0.0103)	(1.1702)				
−0.0382		0.9398		0.0000				
(0.0614)		(0.1569)		(0.5173)				
0.0151								
(0.0036)								
0.2341								
(0.0273)								

presents a scatter plot of estimated random effects overlaid on a set of contour lines, together with summary histograms of the marginal densities. A visual inspection of Figure 3 signifies that there is considerable asymmetry among the estimated random effects. Moreover, the contour level curves appear to follow somewhat satisfactorily the trend of a scattergram, reflecting the adequacy of using a bivariate SN distribution for the random effects.

DOI: 10.1002/sim

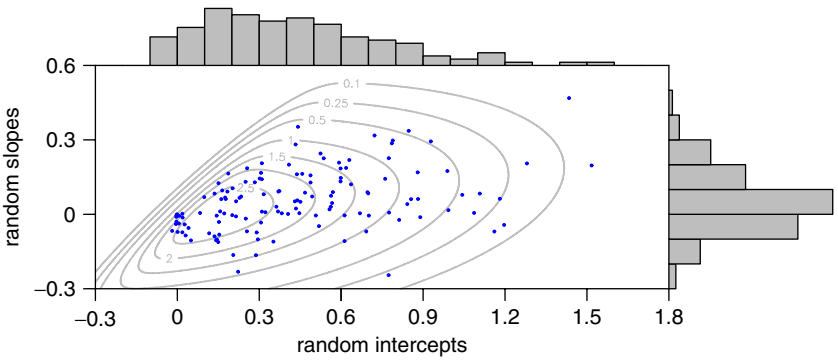

Figure 3. Scatter plot of estimated random effects overlaid on a set of contour lines, together with summary histograms of the marginal densities.

7. SIMULATIONS

The prediction problem for longitudinal data is of great importance in a number of practical applications. As pointed out by Rao [21] and Lee [22], the predictive accuracy of future observations can be taken as an alternative measure of 'fitness' of the data. In this section, we present a simulation study in which attention is focused on comparing the predictive abilities of the NLMM and SNLMM models. In the simulations, we generate 500 Monte Carlo data sets from the model:

$$Y_{ij} = \beta_0 + t_{ij}\beta_1 + w_i\beta_2 + b_i + \varepsilon_{ij} \quad i = 1 \ldots, N, \quad j = 1, \ldots, 6 \tag{21}$$

where $t_{ij} = j$, $w_i = 1$ if $i \leqslant N/2$ and is zero otherwise, $\beta_1 = 2$, $\beta_2 = 1$, $\varepsilon_{ij} \sim N(0, 2^2)$ and $\beta_0 + b_i$ is taken to have the Gamma(2, 1) distribution with density $f(x) = x\exp(-x)$ for yielding a highly right skewed distribution. For each generated data set, model (21) will be fitted twice under the model assumptions outlined in Section 2, with the density of b_i represented by a SN distribution and a normal distribution, respectively.

We use the pseudo-cross-validation approach to assess their predictive performances. This approach of comparing forecasts with the corresponding actual values is in the spirit of cross validation of Stone [23] and Geisser [24]. The technique proceeds as follows: (i) drop out the last three measurements y_{i4}, y_{i5}, y_{i6} on the ith participant; (ii) compute ML estimates using the remaining data as the sample; (iii) prediction of $\mathbf{y}_i = (y_{i4}, y_{i5}, y_{i6})$, denoted by $\hat{\mathbf{y}}_i = (\hat{y}_{i4}, \hat{y}_{i5}, \hat{y}_{i6})$, is made by using formula (16). For each of the 500 generated Monte Carlo data sets, the procedure is repeated across subjects $i = 1, \ldots, N$.

To evaluate prediction accuracies, the quantity of the empirical mean-squared forecast error

$$\text{MSFE} = \frac{1}{3N}\sum_{i=1}^{N}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^{\mathrm{T}}(\hat{\mathbf{y}}_i - \mathbf{y}_i)$$

is calculated as a discrepancy measure. On the basis of this measure, the best predictor corresponds to having the smallest MSFE from a given set of alternatives.

We consider the cases with sample sizes $N = 30, 50, 100, 150$ and comparisons of both predictors based on 500 replications are summarized in Table III. As expected, the numerical results indicated that the SNLMM predictor performs encouragingly well in all cases. In all 2000

DOI: 10.1002/sim

Table III. Comparison of prediction accuracy of NLMM and SNLMM predictors in terms of mean and standard deviation of 500 simulated MSFEs.

	$N = 30$		$N = 50$		$N = 100$		$N = 150$	
Predictor	Mean	SD	Mean	SD	Mean	SD	Mean	SD
NLMM	1.2719	0.2010	1.2920	0.1593	1.2974	0.1116	1.3016	0.0876
SNLMM	1.2184	0.1936	1.2359	0.1516	1.2380	0.1063	1.2413	0.0826

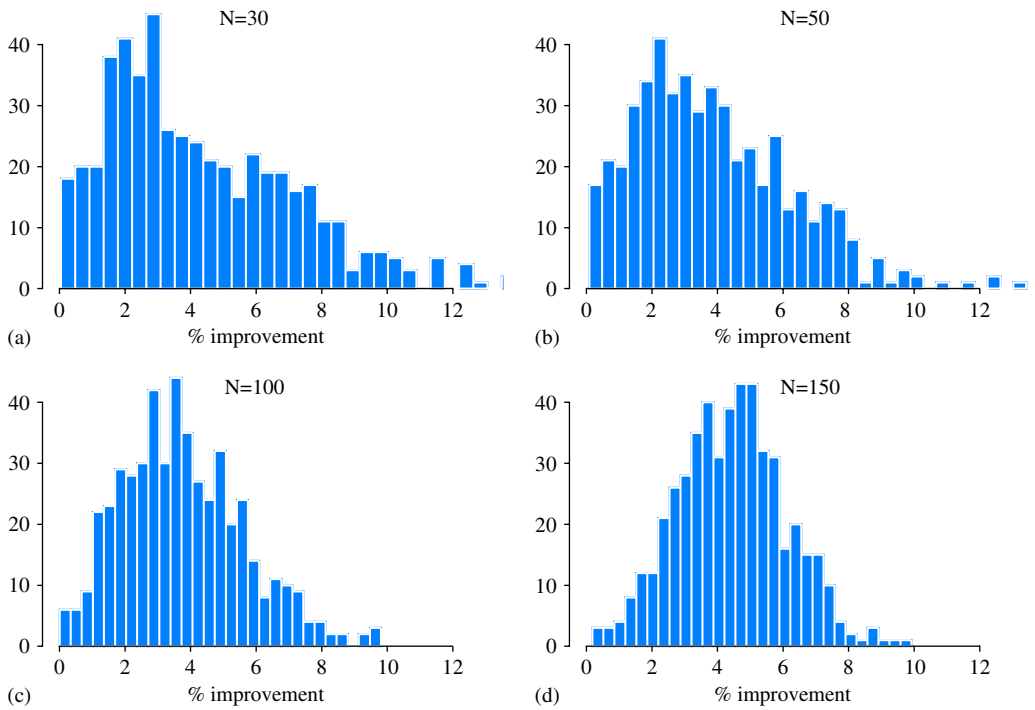

Figure 4. Percentage improvement of the SNLMM predictor over the NLMM predictor for 500 simulated MSFEs: (a) $N = 30$; (b) $N = 50$; (c) $N = 100$; and (d) $N = 150$.

generated simulation data, the SNLMM predictors yield better forecasts (with lower MSFEs) than the classical NLMM predictors. Judging from this, it is important to emphasize that model specification with regard to random effects is noteworthy in prediction. Figure 4 displays the percentage improvement of the SNLMM predictor over the NLMM predictor. It is readily seen that the gains in predictive accuracies obtained from the SNLMM model appear striking.

8. CONCLUSIONS

We have proposed an approach to the analysis of linear mixed models using a multivariate SN distribution for the random effects, which will allow practitioners to fit longitudinal data in

DOI: 10.1002/sim

a wide variety of considerations. We have described a normal-truncated normal hierarchy for the SNLMM model and presented a hybrid ECME-NR algorithm for dealing with ML estimation in a flexible complete data framework. Moreover, the score test as well as prediction procedures considered in this paper are easy to implement once the ML estimates are obtained. Numerical results illustrated in Section 6 indicate that the SNLMM model for the Framingham cholesterol data is evidently more adequate than the conventional NLMM model. The simulation study shows that an appropriate specification of random effects can enhance predictive abilities.

There are a number of possible extensions of the current work. A natural generalization is to robustify the skew-normal distribution using a broader family such as the skew t [25–27] and the skew-elliptical [8, 28] distribution. Another worthwhile task is to develop workable Markov chain Monte Carlo algorithms in a Bayesian framework of hierarchical SNLMM models. Finally, one referee pointed out that it is also of interest to compare SNLMM and t linear mixed models [2, 29, 30] with respect to their robust inferences.

APPENDIX A: THE SCORE FUNCTION AND EMPIRICAL INFORMATION MATRIX

The score vector $\mathbf{s}_{\boldsymbol{\theta}}$ is the vector of the first derivatives of (8) with respect to $\boldsymbol{\theta}=(\boldsymbol{\beta}, \sigma, \boldsymbol{\omega})$, where $\boldsymbol{\omega}=(\text{vech}(\mathbf{F}), \boldsymbol{\rho}, \boldsymbol{\delta})$ with vech($\mathbf{F}$) being the distinct elements of $\mathbf{F}$. Expressions for the elements of $\mathbf{s}_{\boldsymbol{\theta}}$ are

$$\mathbf{s}_{\boldsymbol{\beta}}=\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{Y})}{\partial \boldsymbol{\beta}}=\frac{1}{\sigma^2}\sum_{i=1}^{N}\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i-\frac{1}{\sigma}\sum_{i=1}^{N}\kappa_i\boldsymbol{\xi}_i$$

$$\mathbf{s}_{\sigma}=\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{Y})}{\partial \sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{N}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i-\frac{1}{\sigma}\sum_{i=1}^{N}\kappa_i\eta_i$$

$$[\mathbf{s}_{\boldsymbol{\omega}}]_r=\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{Y})}{\partial \omega_r}=-\frac{1}{2}\sum_{i=1}^{N}\text{tr}(\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir})+\frac{1}{2\sigma^2}\sum_{i=1}^{N}\mathbf{e}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\dot{\boldsymbol{\Lambda}}_{ir}\boldsymbol{\Lambda}_i^{-1}\mathbf{e}_i+\sum_{i=1}^{N}\kappa_i\dot{\eta}_{ir}$$

where κ_i and η_i are as in (7),

$$\boldsymbol{\xi}_i=\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}{\sqrt{1+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}} \quad \dot{\boldsymbol{\Lambda}}_{ir}=\frac{\partial \boldsymbol{\Lambda}_i}{\partial \omega_r} \quad \text{and} \quad \dot{\eta}_{ir}=\frac{\partial \eta_i}{\partial \omega_r}$$

Note that

$$\dot{\boldsymbol{\Lambda}}_{ir}=\begin{cases}\mathbf{Z}_i\dfrac{\partial \mathbf{F}}{\partial f_{ab}}\mathbf{F}\mathbf{Z}_i^{\mathrm{T}}+\mathbf{Z}_i\mathbf{F}\dfrac{\partial \mathbf{F}}{\partial f_{ab}}\mathbf{Z}_i^{\mathrm{T}} & \text{if } \omega_r=f_{ab} \quad (a=1,\ldots,q;\ b=a,a+1,\ldots,q)\\ \dfrac{\partial \mathbf{C}_i}{\partial \rho_a} & \text{if } \omega_r=\rho_a \quad (a=1,\ldots,g)\\ \mathbf{0} & \text{if } \omega_r=\lambda_b \quad (b=1,\ldots,q)\end{cases}$$

and

$$\dot{\eta}_{ir}=\frac{(\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1})\mathbf{e}_i}{\sigma\sqrt{1+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}}-\frac{\eta_i(2\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)}{2(1+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)}$$

DOI: 10.1002/sim

where

$$\dot{\mathbf{d}}_{ir} = \frac{\partial \mathbf{d}_i}{\partial \omega_r} = \begin{cases} \mathbf{Z}_i \dfrac{\partial \mathbf{F}}{\partial f_{ab}} \boldsymbol{\delta} & \text{if } \omega_r = f_{ab} \\ \mathbf{0} & \text{if } \omega_r = \rho_a \\ \mathbf{Z}_i \mathbf{F} \dfrac{\partial \boldsymbol{\delta}}{\partial \lambda_b} & \text{if } \omega_r = \lambda_b \end{cases}$$

and $\dot{\boldsymbol{\Psi}}_{ir} = \partial \boldsymbol{\Psi}_i / \partial \omega_r = \dot{\boldsymbol{\Lambda}}_{ir} - \dot{\mathbf{d}}_{ir}\mathbf{d}_i^{\mathrm{T}} - \mathbf{d}_i \dot{\mathbf{d}}_{ir}^{\mathrm{T}}$. Here, $\partial \mathbf{F} / \partial f_{ab}$ is clearly a binary indicator matrix and the elements of $\partial \mathbf{C}_i / \partial \rho_a$ and $\partial \boldsymbol{\delta} / \partial \lambda_b$ can be obtained straightforwardly.

The observed information $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}$, obtained by the negative of the second derivative of (8), has the following components:

$$\begin{aligned}
\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \frac{1}{\sigma^2} \sum_{i=1}^{N} (\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \mathbf{X}_i + \kappa_i (\eta_i + \kappa_i) \boldsymbol{\xi}_i \boldsymbol{\xi}_i^{\mathrm{T}}) \\
\mathbf{I}_{\boldsymbol{\beta}\sigma} &= \frac{2}{\sigma^3} \sum_{i=1}^{N} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i - \frac{1}{\sigma^2} \sum_{i=1}^{N} \kappa_i (1 - \kappa_i \eta_i - \eta_i^2) \boldsymbol{\xi}_i \\
[\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\omega}}]_r &= \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \dot{\boldsymbol{\Lambda}}_{ir} \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i + \frac{1}{\sigma} \sum_{i=1}^{N} (\dot{\kappa}_{ir} \boldsymbol{\xi}_i + \kappa_i \dot{\boldsymbol{\xi}}_{ir}) \\
\mathbf{I}_{\sigma\sigma} &= -\frac{n}{\sigma^2} + \frac{3}{\sigma^4} \sum_{i=1}^{N} \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i + \frac{1}{\sigma^2} \sum_{i=1}^{N} \kappa_i \eta_i (\eta_i^2 + \kappa_i \eta_i - 2) \\
[\mathbf{I}_{\sigma\boldsymbol{\omega}}]_r &= \frac{1}{\sigma^3} \sum_{i=1}^{N} \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \dot{\boldsymbol{\Lambda}}_{ir} \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i + \frac{1}{\sigma} \sum_{i=1}^{N} (\dot{\kappa}_{ir} \eta_i + \kappa_i \dot{\eta}_{ir}) \\
[\mathbf{I}_{\boldsymbol{\omega}\boldsymbol{\omega}}]_{rs} &= \frac{1}{2} \sum_{i=1}^{N} [\mathrm{tr}(\boldsymbol{\Lambda}_i^{-1} \ddot{\boldsymbol{\Lambda}}_{irs}) - \mathrm{tr}(\boldsymbol{\Lambda}_i^{-1} \dot{\boldsymbol{\Lambda}}_{is} \boldsymbol{\Lambda}_i^{-1} \dot{\boldsymbol{\Lambda}}_{ir})] + \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \dot{\boldsymbol{\Lambda}}_{ir} \boldsymbol{\Lambda}_i^{-1} \dot{\boldsymbol{\Lambda}}_{is} \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i \\
&\quad - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i^{-1} \ddot{\boldsymbol{\Lambda}}_{irs} \boldsymbol{\Lambda}_i^{-1} \mathbf{e}_i - \sum_{i=1}^{N} (\dot{\kappa}_{ir} \dot{\eta}_{ir} + \kappa_i \ddot{\eta}_{irs})
\end{aligned}$$

where

$$\begin{aligned}
\dot{\kappa}_{ir} &= \frac{\partial \kappa_i}{\partial \omega_r} = -\kappa_i (\eta_i + \kappa_i) \dot{\eta}_{ir} \\
\dot{\boldsymbol{\xi}}_{ir} &= \frac{\partial \boldsymbol{\xi}_i}{\partial \omega_r} = -\frac{\mathbf{X}_i^{\mathrm{T}} (\boldsymbol{\Psi}_i^{-1} \dot{\boldsymbol{\Psi}}_{ir} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i - \boldsymbol{\Psi}_i^{-1} \dot{\mathbf{d}}_{ir})}{\sqrt{1 + \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i}} - \frac{\boldsymbol{\xi}_i (2 \dot{\mathbf{d}}_{ir}^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i - \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \dot{\boldsymbol{\Psi}}_{ir} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i)}{2(1 + \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i)} \\
\ddot{\eta}_{irs} &= \frac{\partial^2 \eta_i}{\partial \omega_r \partial \omega_s} \\
&= \frac{A \mathbf{e}_i}{\sigma \sqrt{1 + \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i}} - \frac{(\dot{\mathbf{d}}_{ir}^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} - \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \dot{\boldsymbol{\Psi}}_{ir} \boldsymbol{\Psi}_i^{-1}) \mathbf{e}_i (2 \dot{\mathbf{d}}_{is}^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i - \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \dot{\boldsymbol{\Psi}}_{is} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i)}{2\sigma (1 + \mathbf{d}_i^{\mathrm{T}} \boldsymbol{\Psi}_i^{-1} \mathbf{d}_i)^{3/2}}
\end{aligned}$$

DOI: 10.1002/sim

$$-\frac{\dot{\eta}_{is}(2\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)}{2(1+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)}-\frac{\eta_i B}{2(1+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)}$$

$$-\frac{\eta_i(2\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)(2\dot{\mathbf{d}}_{is}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{is}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)}{2(1+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)^2}$$

with

$$\begin{aligned}A&=\ddot{\mathbf{d}}_{irs}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}-\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{is}\boldsymbol{\Psi}_i^{-1}-\dot{\mathbf{d}}_{is}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{is}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\\&\quad-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\ddot{\boldsymbol{\Psi}}_{irs}\boldsymbol{\Psi}_i^{-1}+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{is}\boldsymbol{\Psi}_i^{-1}\\B&=2(\ddot{\mathbf{d}}_{irs}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i-\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{is}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i+\dot{\mathbf{d}}_{ir}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\mathbf{d}}_{is}-\dot{\mathbf{d}}_{is}^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\\&\quad+\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{is}\boldsymbol{\Psi}_i^{-1}\dot{\boldsymbol{\Psi}}_{ir}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i)-\mathbf{d}_i^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}\ddot{\boldsymbol{\Psi}}_{irs}\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\end{aligned}$$

$$\ddot{\boldsymbol{\Psi}}_{irs}=\ddot{\boldsymbol{\Lambda}}_{irs}-\ddot{\mathbf{d}}_{irs}\mathbf{d}_i^{\mathrm{T}}-\dot{\mathbf{d}}_{ir}\dot{\mathbf{d}}_{is}^{\mathrm{T}}-\dot{\mathbf{d}}_{is}\dot{\mathbf{d}}_{ir}^{\mathrm{T}}-\mathbf{d}_i\ddot{\mathbf{d}}_{irs}^{\mathrm{T}}$$

$$\ddot{\boldsymbol{\Lambda}}_{irs}=\frac{\partial\boldsymbol{\Lambda}_i}{\partial\omega_r\partial\omega_s}=\begin{cases}\dfrac{\partial^2\mathbf{C}_i}{\partial\rho_a\rho_b} & \text{if } \omega_r=\rho_a \text{ and } \omega_s=\rho_b\\ \mathbf{0} & \text{o.w.}\end{cases}$$

and

$$\ddot{\mathbf{d}}_{irs}=\frac{\partial^2\mathbf{d}_i}{\partial\omega_r\partial\omega_s}=\begin{cases}\mathbf{Z}_i\dfrac{\partial\mathbf{F}}{\partial f_{ab}}\dfrac{\partial\boldsymbol{\delta}}{\partial\lambda_c} & \text{if } \omega_r=f_{ab} \text{ and } \omega_s=\lambda_c\\ \mathbf{0} & \text{o.w.}\end{cases}$$

APPENDIX B: PROOFS OF (16) AND (17)

By applying Proposition 4 of [6], the moment generating function of the conditional density in (15) is given by

$$\begin{aligned}M_{\mathbf{y}_i|\mathbf{Y}_i}(\mathbf{t})&=\frac{\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}_{2\cdot1}+\frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t}\right)}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}\int\Phi(\boldsymbol{\upsilon}_i^{(1)\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})+\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}(\mathbf{y}_i-\mathbf{x}_i\boldsymbol{\beta}))\\&\quad\times(2\pi)^{-\nu/2}|\boldsymbol{\Omega}_{22\cdot1}|^{-1/2}\exp\{-\tfrac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_{2\cdot1}-\boldsymbol{\Omega}_{22\cdot1}\mathbf{t})^{\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_{2\cdot1}-\boldsymbol{\Omega}_{22\cdot1}\mathbf{t})\}\mathrm{d}\mathbf{y}_i\\&=\frac{\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}_{2\cdot1}+\frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t}\right)}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}))}\Phi\left(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta})+\frac{\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t}}{\sqrt{1+\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}}\right)\quad(\mathbf{t}\in\mathbb{R}^{\nu})\end{aligned}$$

DOI: 10.1002/sim

Differentiating $M_{\mathbf{y}_i|\mathbf{Y}_i}(\mathbf{t})$ with respect to $\mathbf{t}$ we get

$$M'_{\mathbf{y}_i|\mathbf{Y}_i}(\mathbf{t}) = \frac{(\boldsymbol{\mu}_{2\cdot1} + \boldsymbol{\Omega}_{22\cdot1}\mathbf{t})\exp(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}_{2\cdot1} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t})}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}))}\Phi\left(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \frac{\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t}}{\sqrt{1 + \boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}}\right)$$
$$+ \frac{\exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}_{2\cdot1} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t}\right)}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}))}\phi\left(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \frac{\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\mathbf{t}}{\sqrt{1 + \boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}}\right)$$
$$\times \frac{\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}{\sqrt{1 + \boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}}$$

Substituting $\mathbf{t} = \mathbf{0}$ into $M'_{\mathbf{y}_i|\mathbf{Y}_i}(\mathbf{t})$, we complete the proof of (16). It therefore suffices to verify that

$$E(\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}|\mathbf{Y}_i, \boldsymbol{\theta}) = \frac{\partial}{\partial\mathbf{t}^{\mathrm{T}}}M'_{\mathbf{y}_i|\mathbf{Y}_i}(\mathbf{t})\Big|_{\mathbf{t}=\mathbf{0}}$$
$$= \boldsymbol{\Omega}_{22\cdot1} + \boldsymbol{\mu}_{2\cdot1}\boldsymbol{\mu}_{2\cdot1}^{\mathrm{T}} + \frac{\phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}))}{\Phi(\tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}))}\left[\frac{\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}\boldsymbol{\mu}_{2\cdot1}^{\mathrm{T}}}{\sqrt{1 + \boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}}\right.$$
$$\left. + \frac{\boldsymbol{\mu}_{2\cdot1}\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}}{\sqrt{1 + \boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}} - \tilde{\boldsymbol{\upsilon}}_i^{\mathrm{T}}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\frac{\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}\boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}}{1 + \boldsymbol{\upsilon}_i^{(2)\mathrm{T}}\boldsymbol{\Omega}_{22\cdot1}\boldsymbol{\upsilon}_i^{(2)}}\right]$$

Using the fact that $\mathrm{cov}(\mathbf{y}_i|\mathbf{Y}_i, \boldsymbol{\theta}) = E(\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}|\mathbf{Y}_i, \boldsymbol{\theta}) - E(\mathbf{y}_i|\mathbf{Y}_i, \boldsymbol{\theta})E(\mathbf{y}_i|\mathbf{Y}_i, \boldsymbol{\theta})^{\mathrm{T}}$, the result (17) is then proved.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Arellano-Valle for providing the Framingham cholesterol data at an early stage of this work. We are also grateful to the editor and two anonymous referees for their valuable comments and suggestions, which led to substantial improvements in the presentation of this work. This research was partly supported by the National Science Council of Taiwan (grant no. NSC95-2118-M-005-001-MY2).

REFERENCES

1. Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
2. Pinheiro JC, Liu CH, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 2001; **10**:249–276.
3. Verberk G, Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 1996; **91**:217–221.
4. Lin TI, Lee JC. On modelling data from degradation sample paths over time. *Australian and New Zealand Journal of Statistics* 2003; **45**:257–270.
5. Lee JC, Lin TI, Lee KJ, Hsu YL. Bayesian analysis of Box–Cox transformed linear mixed models with ARMA(p, q) dependence. *Journal of Statistical Planning and Inference* 2005; **133**:435–451.

DOI: 10.1002/sim

6. Azzalini A, Dalla Valle A. The multivariate skew-normal distribution. *Biometrika* 1996; **83**:715–726.
7. Azzalini A, Capitaino A. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:579–602.
8. Sahu SK, Dey DK, Branco MD. A new class of multivariate skew distributions with application to Bayesian regression models. *Canadian Journal of Statistics* 2003; **31**:129–150.
9. Arellano-Valle RB, Genton MG. On fundamental skew distributions. *Journal of Multivariate Analysis* 2005; **96**:93–116.
10. González-Farias G, Domínguez-Molina A, Gupta AK. Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference* 2004; **126**:521–534.
11. Liseo B, Loperfido N. A Bayesian interpretation of the multivariate skew-normal distribution. *Statistics and Probability Letters* 2003; **61**:395–401.
12. Azzalini A. The skew-normal distribution and related multivariate families (with discussion). *Scandinavian Journal of Statistics* 2005; **32**:159–200.
13. Arellano-Valle RB, Azzalini A. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* 2006; **33**:561–574.
14. Arellano-Valle RB, Bolfarine H, Lachos VH. Skew-normal linear mixed models. *Journal of Data Science* 2005; **3**:415–438.
15. Genton MG. *Skew-elliptical Distributions and Their Applications*. Chapman & Hall: New York, 2004.
16. Thisted RA. *Elements of Statistical Computing*: *Numerical Computation*. Chapman & Hall: New York, 1988.
17. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.
18. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993; **80**:267–278.
19. Liu CH, Rubin DB. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* 1994; **81**:633–648.
20. Zhang D, Davidian M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* 2001; **57**:795–802.
21. Rao CR. Prediction of future observations in growth curve models. *Statistical Science* 1987; **2**:434–471.
22. Lee JC. Prediction and estimation of growth curve with special covariance structures. *Journal of the American Statistical Association* 1988; **83**:432–440.
23. Stone M. Cross-validatory choice and assessment of statistical prediction (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1974; **36**:111–147.
24. Geisser S. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 1975; **70**:320–328.
25. Jones MC, Faddy MJ. A skew extension of the *t*-distribution, with applications. *Journal of the Royal Statistical Society*, *Series B* 2003; **65**:159–174.
26. Azzalini A, Capitaino A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution. *Journal of the Royal Statistical Society*, *Series B* 2003; **65**:367–389.
27. Lin TI, Lee JC, Hsieh WJ. Robust mixture modeling using the skew *t* distribution. *Statistics and Computing* 2007; **17**:81–92.
28. Branco MD, Dey DK. A general class of multivariate skew elliptical distributions. *Journal of Multivariate Analysis* 2001; **79**:99–113.
29. Lin TI, Lee JC. A robust approach to *t* linear mixed models applied to multiple sclerosis data. *Statistics in Medicine* 2006; **25**:1397–1412.
30. Lin TI, Lee JC. Bayesian analysis of hierarchical linear mixed modeling using the multivariate *t* distribution. *Journal of Statistical Planning and Inference* 2007; **137**:484–495.

DOI: 10.1002/sim